

CLAIMS

1. An index term extraction device, comprising:

input means for inputting a document-to-be-surveyed, documents-to-be-compared to be compared with said document-to-be-surveyed, and source-documents-for-selection to become the selection source of similar documents that are similar to said document-to-be-surveyed;

index term extraction means for extracting index terms from said document-to-be-surveyed;

first appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

similar documents selecting means for selecting said similar documents from said source-documents-for-selection based on data of said document-to-be-surveyed;

second appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

output means for outputting each index term and positioning data thereof, based on the combination of the calculated function value of the appearance frequency in said documents-to-be-compared and the calculated function value of the appearance frequency in said similar documents, regarding each index term.

2. The index term extraction device according to claim 1, wherein said documents-to-be-compared are used as said source-documents-for-selection.

3. The index term extraction device according to claim 1 or claim 2, wherein said similar documents selecting means calculates, with respect to each document of said document-to-be-surveyed and said source-documents-for-selection, a vector having as its component a function value of an appearance frequency in each document of each index term contained in each document, or a function value of an appearance frequency in said source-documents-for-selection of each index term contained in each document; and selects from said source-documents-for-selection documents having a vector of a high degree of similarity to said vector calculated with respect to said document-to-be-surveyed, and makes the selected documents similar documents.

4. The index term extraction device according to any one of claims 1 to 3, wherein said output means outputs, based on the results of the respective calculation means, an index term of a first group having a low appearance frequency in said documents-to-be-compared and in said similar documents, an index term of a second group having a higher appearance

frequency in said documents-to-be-compared in comparison to the index term of said first group, and an index term of a third group having a higher appearance frequency in said similar documents in comparison to the index term of said first group.

5. The index term extraction device according to any one of claims 1 to 3, wherein said output means outputs, based on the results of the respective calculation means, an index term of a third group having a lower appearance frequency in said documents-to-be-compared in comparison to an index term of a fourth group having a high appearance frequency in said documents-to-be-compared and in said similar documents, an index term of a second group having a lower appearance frequency in said similar documents in comparison to the index term of said fourth group, and an index term of a first group having a lower appearance frequency in said similar documents in comparison to the index term of said third group and further having a lower appearance frequency in said documents-to-be-compared in comparison to the index term of said second group.

6. An index term extraction device, comprising:

input means for inputting a document-to-be-surveyed, documents-to-be-compared to be compared with said document-to-

be-surveyed, and similar documents that are similar to said document-to-be-surveyed;

index term extraction means for extracting index terms from said document-to-be-surveyed;

first appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

second appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

output means for outputting, based on the results of the respective calculation means,

an index term of a first group having a low appearance frequency in said documents-to-be-compared and in said similar documents,

an index term of a second group having a higher appearance frequency in said documents-to-be-compared in comparison to the index term of said first group, and

an index term of a third group having a higher appearance frequency in said similar documents in comparison to the index term of said first group.

7. An index term extraction device, comprising:

input means for inputting a document-to-be-surveyed,

documents-to-be-compared to be compared with said document-to-be-surveyed, and similar documents that are similar to said document-to-be-surveyed;

index term extraction means for extracting index terms from said document-to-be-surveyed;

first appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

second appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

output means for outputting, based on the results of the respective calculation means,

an index term of a third group having a lower appearance frequency in said documents-to-be-compared in comparison to an index term of a fourth group having a high appearance frequency in said documents-to-be-compared and in said similar documents,

an index term of a second group having a lower appearance frequency in said similar documents in comparison to the index term of said fourth group, and

an index term of a first group having a lower appearance frequency in said similar documents in comparison to the index term of said third group and further having a lower appearance

frequency in said documents-to-be-compared in comparison to the index term of said second group.

8. The index term extraction device according to any one of claims 1 to 7, wherein the function value of the appearance frequency in said documents-to-be-compared or said similar documents is a logarithm of a value obtained by multiplying the total number of documents of said documents-to-be-compared or said similar documents to the reciprocal of said appearance frequency.

9. The index term extraction device according to any one of claims 1 to 8, wherein said output means disposes and outputs each index term by taking the function value of the appearance frequency in said documents-to-be-compared as a first axis of a coordinate system and taking the function value of the appearance frequency in said similar documents as a second axis of said coordinate system.

10. The index term extraction device according to any one of claims 4 to 8, wherein said output means respectively lists and outputs the index term of said first group, the index term of said second group, and the index term of said third group.

11. The index term extraction device according to any one of claims 4 to 8, wherein said output means automatically creates

and outputs supporting documentation of said document-to-be-surveyed through the use of the index term of said first group, the index term of said second group, and the index term of said third group.

12. The index term extraction device according to any one of claims 1 to 8,

wherein each of said similar documents is included in said documents-to-be-compared,

wherein said output means disposes and outputs each index term by further transforming the function value of the appearance frequency in said documents-to-be-compared and taking the same as a first axis of a coordinate system and taking the function value of the appearance frequency in said similar documents as a second axis of said coordinate system, and

wherein said transformation is conducted such that a boundary line of an existable area of said index terms on said coordinate system, based on said similar documents being a subset of said documents-to-be-compared, approaches vertical line of said first axis.

13. The index term extraction device according to claim 12, wherein said transformation is given according to the function with the appearance frequency in said similar documents.

14. The index term extraction device according to any one of claims 1 to 13,

further comprising term frequency calculation means for calculating an appearance frequency, in said document-to-be-surveyed, of each index term in said document-to-be-surveyed,

wherein said output means reflects and outputs the appearance frequency, in said document-to-be-surveyed, of each index term in said document-to-be-surveyed.

15. The index term extraction device according to any one of claims 1 to 8, wherein, when said output means, for each index term, takes the function value of the appearance frequency in said documents-to-be-compared as a first axis of a coordinate system and takes the function value of the appearance frequency in said similar documents as a second axis of said coordinate system, said output means disposes each index term so as to further approach a reference point that is the closest to said index term among a plurality of reference points on said coordinate system and outputs each index term on said coordinate system.

16. The index term extraction device according to any one of claims 1 to 8, further comprising:

reference point setting means for setting coordinates of a plurality of reference points on a coordinate system;

means for updating a prescribed number of times the

coordinate data of a reference point that is closest to said index term among said plurality of reference points so as to further approach said index term when, for each index term, the function value of the appearance frequency in said documents-to-be-compared is taken as a first axis of the coordinate system and the function value of the appearance frequency in said similar documents is taken as a second axis of said coordinate system; and

coordinate calculation means for calculating coordinates for disposing said index term based on said updated reference point,

wherein said output means disposes and outputs each index term on said coordinate system based on the coordinates calculated by said coordinate calculation means.

17. An index term extraction method, comprising:

an input step for inputting a document-to-be-surveyed, documents-to-be-compared to be compared with said document-to-be-surveyed, and source-documents-for-selection to become the selection source of similar documents that are similar to said document-to-be-surveyed;

an index term extraction step for extracting index terms from said document-to-be-surveyed;

a first appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-

compared;

similar documents selecting step for selecting said similar documents from said source-documents-for-selection based on data of said document-to-be-surveyed;

a second appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

an output step for outputting each index term and positioning data thereof based on the combination of the calculated function value of the appearance frequency in said documents-to-be-compared and the calculated function value of the appearance frequency in said similar documents, regarding each index term.

18. An index term extraction method, comprising:

an input step for inputting a document-to-be-surveyed, documents-to-be-compared to be compared with said document-to-be-surveyed, and similar documents that are similar to said document-to-be-surveyed;

an index term extraction step for extracting index terms from said document-to-be-surveyed;

a first appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

• • " "

a second appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

an output step for outputting, based on the results of the respective calculation steps,
an index term of a first group having a low appearance frequency in said documents-to-be-compared and in said similar documents,
an index term of a second group having a higher appearance frequency in said documents-to-be-compared in comparison to the index term of said first group, and
an index term of a third group having a higher appearance frequency in said similar documents in comparison to the index term of said first group.

19. An index term extraction program for causing a computer to execute:

an input step for inputting a document-to-be-surveyed, documents-to-be-compared to be compared with said document-to-be-surveyed, and source-documents-for-selection to become the selection source of similar documents that are similar to said document-to-be-surveyed;

an index term extraction step for extracting index terms from said document-to-be-surveyed;

a first appearance frequency calculation step for

calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

similar documents selecting step for selecting said similar documents from said source-documents-for-selection based on data of said document-to-be-surveyed;

a second appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

an output step for outputting each index term and positioning data thereof based on the combination of the calculated function value of the appearance frequency in said documents-to-be-compared and the calculated function value of the appearance frequency in said similar documents, regarding each index term.

20. An index term extraction program for causing a computer to execute:

an input step for inputting a document-to-be-surveyed, documents-to-be-compared to be compared with said document-to-be-surveyed, and similar documents that are similar to said document-to-be-surveyed;

an index term extraction step for extracting index terms from said document-to-be-surveyed;

a first appearance frequency calculation step for

calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

a second appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said similar documents; and

an output step for outputting, based on the results of the respective calculation steps,

an index term of a first group having a low appearance frequency in said documents-to-be-compared and in said similar documents,

an index term of a second group having a higher appearance frequency in said documents-to-be-compared in comparison to the index term of said first group, and

an index term of a third group having a higher appearance frequency in said similar documents in comparison to the index term of said first group.

21. A character representative diagram of a document-to-be-surveyed, wherein, for each index term in the document-to-be-surveyed,

a function value of an appearance frequency in documents-to-be-compared to be compared with said document-to-be-surveyed is taken as a first axis of a coordinate system, and a function value of an appearance frequency in similar

documents that are similar to said document-to-be-surveyed is taken as a second axis of said coordinate system.

22. A character representative diagram of a document-to-be-surveyed having disposed therein index terms in the document-to-be-surveyed, wherein

an index term of a first group having a low appearance frequency in documents-to-be-compared to be compared with said document-to-be-surveyed and in similar documents that are similar to said document-to-be-surveyed is disposed in a first area,

an index term of a second group having a higher appearance frequency in said documents-to-be-compared in comparison to the index term of said first group is disposed in a second area, and

an index term of a third group having a higher appearance frequency in said similar documents in comparison to the index term of said first group is disposed in a third area.

23. A character representative diagram of a document-to-be-surveyed having disposed therein index terms in the document-to-be-surveyed, wherein

an index term of a third group having a lower appearance frequency in documents-to-be-compared to be compared with said document-to-be-surveyed in comparison to an index term of a fourth group having a high appearance frequency in said

documents-to-be-compared and in similar documents that are similar to said document-to-be-surveyed is disposed in a third area,

an index term of a second group having a lower appearance frequency in said similar documents in comparison to the index term of said fourth group is disposed in a second area, and

an index term of a first group having a lower appearance frequency in said similar documents in comparison to the index term of said third group and further having a lower appearance frequency in said documents-to-be-compared in comparison to the index term of said second group is disposed in a first area.

24. A document characteristic analysis device, comprising:

input means for inputting a document-group-to-be-surveyed including a plurality of documents-to-be-surveyed, documents-to-be-compared to be compared with each document-to-be-surveyed, and related documents having a common attribute with said document-group-to-be-surveyed;

index term extraction means for extracting index terms in each document-to-be-surveyed;

third appearance frequency calculation means for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

fourth appearance frequency calculation means for

calculating a function value of an appearance frequency of each of said extracted index terms in said related documents;

central point calculation means for calculating a central point in each document-to-be-surveyed based on the combination of the calculated function value of the appearance frequency in said documents-to-be-compared and the calculated function value of the appearance frequency in said related documents, regarding each index term; and

output means for outputting data of said central point in each document-to-be-surveyed.

25. The document characteristic analysis device according to claim 24, wherein the calculation of said central point in each document-to-be-surveyed is conducted by calculating the weighted average of the index term coordinates, which is an average value obtained by performing weighting to the coordinate value of each index term based on the function value of the appearance frequency in said documents-to-be-compared and the function value of the appearance frequency in said related documents, regarding each index term, with the ratio of term frequency value of each index term in relation to term frequency value total in said documents.

26. The document characteristic analysis device according to claim 24 or claim 25, wherein data of said central point is output by extracting documents each having high similarity to

• • • •
said document-group-to-be-surveyed and documents each having low similarity to said document-group-to-be-surveyed, among said document-group-to-be-surveyed.

27. A document characteristic analysis method, comprising:

 an input step for inputting a document-group-to-be-surveyed including a plurality of documents-to-be-surveyed, documents-to-be-compared to be compared with each document-to-be-surveyed, and related documents having a common attribute with said document-group-to-be-surveyed;

 an index term extraction step for extracting index terms in each document-to-be-surveyed;

 a third appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

 a fourth appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said related documents;

 central point calculation step for calculating a central point in each document-to-be-surveyed based on the combination of the calculated function value of the appearance frequency in said documents-to-be-compared and the calculated function value of the appearance frequency in said related documents, regarding each index term; and

 an output step for outputting data of said central point

in each document-to-be-surveyed.

28. A document characteristic analysis program for causing a computer to execute:

an input step for inputting a document-group-to-be-surveyed including a plurality of documents-to-be-surveyed, documents-to-be-compared to be compared with each document-to-be-surveyed, and related documents having a common attribute with said document-group-to-be-surveyed;

an index term extraction step for extracting index terms in each document-to-be-surveyed;

a third appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said documents-to-be-compared;

a fourth appearance frequency calculation step for calculating a function value of an appearance frequency of each of said extracted index terms in said related documents;

central point calculation step for calculating a central point in each document-to-be-surveyed based on the combination of the calculated function value of the appearance frequency in said documents-to-be-compared and the calculated function value of the appearance frequency in said related documents, regarding each index term; and

an output step for outputting data of said central point in each document-to-be-surveyed.

29. A document characteristic representative diagram of documents-to-be-surveyed, regarding each of a plurality of documents included in the documents-to-be-surveyed, taking positioning with respect to documents-to-be-compared to be compared with each document-to-be-surveyed as a first axis of a coordinate system and taking positioning with respect to related documents having a common attribute with said documents-to-be-surveyed as a second axis of said coordinate system, wherein

a coordinate value of each of said documents-to-be-surveyed on said coordinate system is set to be a central point, in each document-to-be-surveyed, of index term coordinate values each having as component thereof a function value of an appearance frequency in said documents-to-be-compared of each index term and a function value of an appearance frequency in said related documents of each index term.